

# Data Engineering Project -2

## Azure Data Monitoring For Tokyo Olympics Data

### Team:

Aathirainathan P  
Ambati Sesha Sai Sahithya  
Eswara Venkata Sai Raja

**Date:** 19-12-2024

### 1. Project Overview:

The Data Quality and Monitoring project leverages Azure's data engineering tools to build a system that ensures high-quality, trustworthy datasets. By integrating Azure Data Factory (ADF) and Azure Databricks, the project monitors, validates, and reports on 10 data quality metrics for the Tokyo Olympics dataset. This ensures robust workflows, enabling seamless analytics and data-driven decision-making.

### 2. Project Statement:

This project focuses on building an automated and scalable data pipeline to:

1. Ingest data from multiple sources (e.g., GitHub) into Azure Data Lake Storage.
2. Perform data quality checks and validations to identify and rectify inconsistencies.
3. Store quality metrics for reporting and decision-making.
4. Enable proactive monitoring with alerting mechanisms for data quality issues.

### 3. Data Overview:

The dataset for this project contains data from the Tokyo Olympics, including details about athletes, teams, coaches, gender participation, and medal counts. Key files include:

- *\*Athletes.csv\**: Athlete details like Name, NOC (National Olympic Committee), and Discipline.
- *\*Coaches.csv\**: Coaches' details and associated events.
- *\*EntriesGender.csv\**: Gender-based participation metrics for various events.
- *\*Medals.csv\**: Medal counts and rankings by country.
- *\*Teams.csv\**: Team event participation information.

Schema Details:

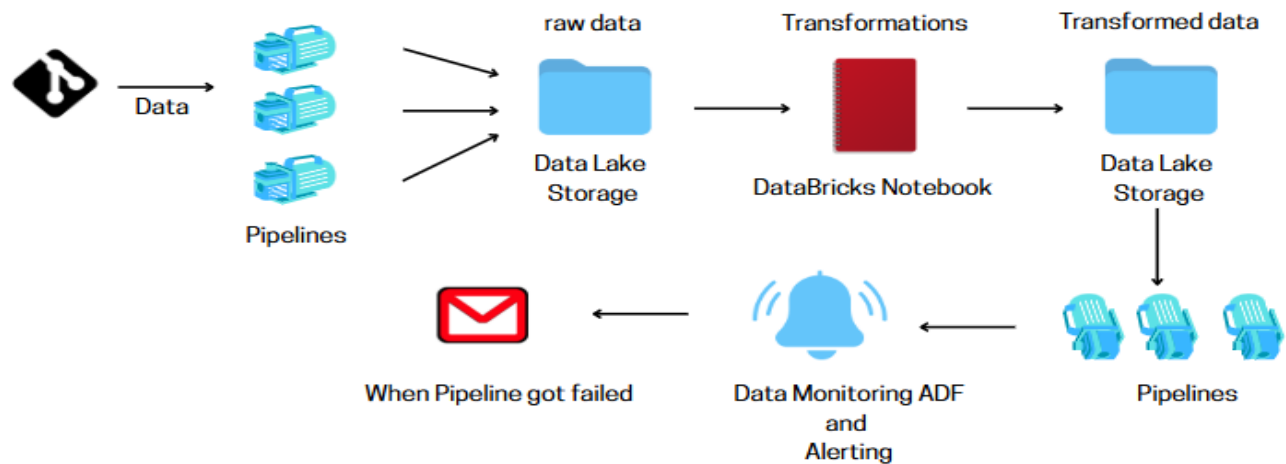
Column Name	Description
Athlete ID	Unique identifier for each athlete.
Name	Name of the athlete.
NOC	National Olympic Committee associated with the athlete.
Discipline	Sports discipline in which the athlete participates.
Coach ID	Unique identifier for each coach.
Coach Name	Name of the coach.
Event	Event associated with the coach or athlete.
Gender	Gender of participants (e.g., Male, Female).
Team	Team associated with the athlete or coach.
Medal	Medal won by the athlete (e.g., Gold, Silver, Bronze).
Rank	Rank of the team or athlete in the event.
Total Participants	Total number of participants in the event.
Category	Category of the event (e.g., Individual, Team).
Region	Geographical region of the participant's NOC (e.g., Asia, Europe).
Country	Country of the athlete or coach.
Event ID	Unique identifier for each event.
Participation Count	Number of participants in each gender category.
Null Count	Count of null or missing values in specific columns.
Duplicate Count	Count of duplicate records identified during quality checks.
Outliers Detected	Number of records with outlier values for specific numeric fields.
Negative Values Found	Number of records with invalid negative values (e.g., negative medal counts).
Quality Metric	Consolidated quality metrics for nulls, duplicates, and validation errors.
Pipeline Status	Status of the pipeline (e.g., Success, Failed).
Processed Date	Date when the data was processed.

## Data Zones

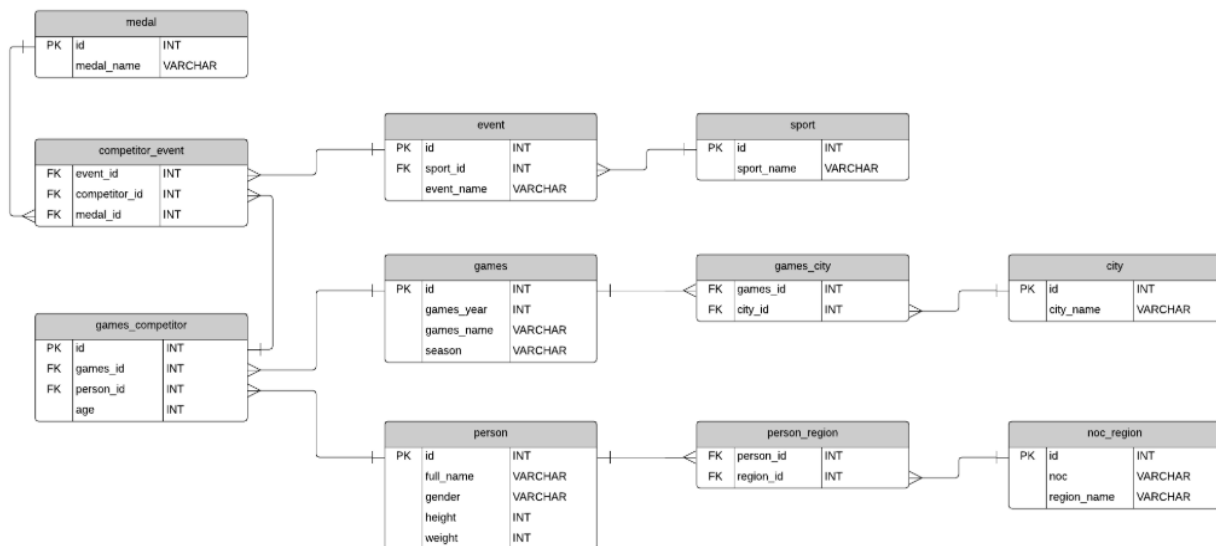
The dataset undergoes the following transformations through the pipeline:

1. **Bronze Zone:** Stores the raw data ingested from Azure Blob Storage.
2. **Silver Zone:** Contains cleaned and standardized data after transformations (e.g., type casting, null handling).
3. **Gold Zone:** Contains aggregated data optimized for analytics and visualization.

## 4. Architectural Diagram:



## 5. ER Diagram:



## 6.How it Works:

1. **\*Data Ingestion\*:**
  - Raw CSV files from GitHub are ingested into Azure Data Lake using ADF pipelines.
2. **\*Data Quality Checks\*:**
  - Databricks notebooks validate the data with 10 quality checks (e.g., null checks, duplicate detection, range validation).
3. **\*Metrics Storage\*:**
  - Quality metrics are stored in CSV format in the Azure Data Lake for reporting.
4. **\*Alerts\*:**
  - Azure Monitor triggers alerts for pipeline failures or quality anomalies.
5. **\*Reporting\*:**
  - Power BI connects to Azure Data Lake to visualize metrics and trends.

## 7. Execution Overview:

### *Step 1: Ingestion (Bronze Zone)*

- The raw data file (Global\_Superstore.csv) is uploaded to **Azure Blob Storage**.
- Azure Data Factory (ADF) pipelines copy the file into **Azure Data Lake Storage Gen2 (ADLS)**.
- The ingestion tasks involve mounting the data in **Databricks** for further processing.

### *Step 2: Transformation (Silver Zone)*

- Using Databricks Notebooks, the raw data is cleaned and transformed into structured data:
  - Data types are cast appropriately (e.g., sales as float, quantity as int).
  - Columns like Profit Margin and High Discount are calculated.
  - Null values and invalid rows are removed.

### *Step 3: Analytical Zone (Gold Zone)*

- The cleaned and aggregated data in the Silver Zone is used to:
  - Compute sales and profit by region, category, and year.
  - Identify high-profit orders and discount trends.
- The processed data is used in Delta format for visualizations and insights.

## **Pipeline Overview:**

1. **\*Ingestion Pipeline\*:**
  - Extracts data from GitHub and loads it into the Raw Zone.
2. **\*Transformation Pipeline\*:**
  - Executes Databricks notebooks for quality checks.
  - Processes data into the Processed Zone.
3. **\*Metrics Pipeline\*:**
  - Consolidates and stores quality metrics in the Metrics Zone.
4. **\*Monitoring Pipeline\*:**
  - Tracks pipeline performance and triggers alerts for anomalies.

## **8. Resources Used for the project:**

- **\*Azure Data Factory\*:** For orchestration of data pipelines.
- **\*Azure Data Lake Storage\*:** To store raw, processed, and metrics data.
- **\*Azure Databricks\*:** For data quality checks and advanced transformations.
- **\*Azure Monitor\*:** For pipeline and quality monitoring with alerts.
- **\*Power BI\*:** For reporting and visualization of quality metrics.
- **\*GitHub\*:** Source repository for raw CSV files.

## **9. Project Requirements:**

- **\*Azure Subscription\*:** with access to ADF, Data Lake Storage, and Databricks.
- **\*Data Sources\*:**
  - Tokyo Olympics dataset in CSV format.
- **\*Tools and Software\*:**
  - Power BI Desktop for report creation.
- **\*Permissions\*:**
  - Access to Azure resources (storage, compute, and monitoring).

10. Tasks Performed:

Creating an Data Lake storage Account:

Home >

tokyoolympicdata9999

Storage account

Search

Upload

Open in Explorer

Delete

Move

Refresh

Open in mobile

CLI / PS

Feedback

Overview

Activity log

Tags

Diagnose and solve problems

Access Control (IAM)

Data migration

Events

Storage browser

Partner solutions

Data storage

Security + networking

Data management

Storage tasks (preview)

Redundancy

Data protection

Blob inventory

Essentials

Resource group (move)

Location

Subscription (move)

Subscription ID

Disk state

Tags (edit)

Performance

Replication

Account kind

Provisioning state

Created

Properties

Monitoring

Capabilities (5)

Recommendations (0)

Tutorials

Tools + SDKs

Data Lake Storage

Hierarchical namespace

Default access tier

Blob anonymous access

Blob soft delete

Container soft delete

Versioning

Change feed

Enabled

Hot

Disabled

Enabled (7 days)

Enabled (7 days)

Disabled

Disabled

Security

Require secure transfer for REST API operations

Storage account key access

Minimum TLS version

Infrastructure encryption

Enabled

Enabled

Version 1.2

Disabled

Networking

Creating container and directories in delta lake storage account:

Home > tokyoolympicdata9999 | Containers >

tokyo-olympic-data

Container

Search

Upload

Add Directory

Refresh

Rename

Delete

Change tier

Acquire lease

Break lease

Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Authentication method: Access key (Switch to Microsoft Entra user account)

Location: tokyo-olympic-data


Search blobs by prefix (case-sensitive)

Show deleted objects

	Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
<input type="checkbox"/>	quality-metrics	12/19/2024, 12:09:31...					- ***
<input type="checkbox"/>	raw-data	12/18/2024, 11:38:01...					- ***
<input type="checkbox"/>	transformed-data	12/18/2024, 11:38:10...					- ***

Creating ADF Account:

[Home](#) >

TokyoOlympicADF9999

✦ ☆ ⋮

✕

Data factory (V2)

Search

◊ ◀

🗑️ Delete

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Settings

Getting started

Monitoring

Automation

Help

Essentials

Resource group (move) : rg-azuser2356\_mmllocal-eylrk

Status : Succeeded

Location : West US 2


Subscription (move) : MML Learners

Subscription ID : 2a3c6418-97b9-4d96-a24b-2c2d7633d375

Type : Data factory (V2)

Getting started : [Quick start](#)

JSON View



Azure Data Factory Studio

Launch studio

Quick Starts

Tutorials

Template Gallery

Training Modules

Uploading CSV Files into github to access through pipeline:

revs-96 / Trail-DE-Project-1-

🔍 Type [Z] to search

👤 + ⌂ 🔍 📧 👤

<> Code

🕒 Issues

🔄 Pull requests

⚙️ Actions

📁 Projects

📖 Wiki

🛡️ Security

📊 Insights

⚙️ Settings

Files

main

Go to file

Dataset

Athletes.csv

Coaches.csv

EntriesGender.csv

Medals.csv

Teams.csv

Trail-DE-Project-1- / Dataset /

Add file ⌵ ⋮

revs-96 Add files via upload 9d88d30 · 5 days ago 🕒 History

Name	Last commit message	Last commit date
..		
Athletes.csv	Add files via upload	5 days ago
Coaches.csv	Add files via upload	5 days ago
EntriesGender.csv	Add files via upload	5 days ago
Medals.csv	Add files via upload	5 days ago
Teams.csv	Add files via upload	5 days ago

## Ingesting data from github(Source) to Data lake storage(sink):

### INGESTION PIPELINE

Microsoft Azure | Data Factory | TokyoOlympicADF9999

Search factory and documentation

Factory Resources

- Pipelines (4)
  - Data\_Quality\_Checks
  - Final\_Pipeline
  - ingestion
  - Store\_Quality\_Metrics
- Change Data Capture (preview) (0)
- Datasets (11)
  - Athletes
  - AthletesData
  - AthletesSink
  - Coaches
  - CoachesSink
  - EntriesGender
  - EntriesGenderSink
  - Medals
  - MedalsSink

ingestion | Athletes

Validate | Validate copy runtime | Debug | Add trigger

Copy data | Copy data | Copy data | Copy data | Copy data

Athletes | Coaches | EntriesGender | Medals | Teams

General | Source | Sink | Mapping | Settings | User properties

Source dataset \* | Athletes | Open | New | Preview data | Learn more

Request method \* | GET

Additional headers

Request body

### Source:

Microsoft Azure | Data Factory | TokyoOlympicADF9999

Search factory and documentation

Factory Resources

- Pipelines (4)
  - Data\_Quality\_Checks
  - Final\_Pipeline
  - ingestion
  - Store\_Quality\_Metrics
- Change Data Capture (preview) (0)
- Datasets (11)
  - Athletes
  - AthletesData
  - AthletesSink
  - Coaches
  - CoachesSink
  - EntriesGender
  - EntriesGenderSink
  - Medals
  - MedalsSink
  - Teams

ingestion | Athletes

Delimited Text | Athletes

CSV

Connection | Schema | Parameters

Linked service \* | HttpAthletes | Test connection

Base URL | https://raw.githubusercontent.com/revs-96/Trail-DE-Project-1-/main/Dataset/Athletes.csv

Relative URL

Compression type | No compression

Column delimiter | Comma (,)

Row delimiter | Default (\r\n, or \n)

Encoding | Default(UTF-8)

Quote character | Double quote (")

Edit linked service

HTTP | Learn more

Name \* | HttpAthletes

Description

Connect via integration runtime \* | AutoResolveIntegrationRuntime

Base URL \* | https://raw.githubusercontent.com/revs-96/Trail-DE-Project-1-/main/Dataset/Athletes.csv

Information will be sent to the URL specified. Please ensure you trust the URL entered.

Server certificate validation | Enable | Disable

Authentication type \* | Anonymous

Auth headers | + New

Annotations

Save | Cancel | Test connection



Sink:

Microsoft Azure

Data FactoryTokyoOlympicADF9999

Search factory and documentation

azuser2356\_mml.local@techademy.com  
TECHADEMY LEARNING SOLUTIONS PRIVATE LIMITED

>>

Data Factory

Validate all

Publish all

Preview experienceOff

Factory Resources

Filter resources by name

Pipelines4

- Data\_Quality\_Checks
- Final\_Pipeline
- ingestion
- Store\_Quality\_Metrics

Change Data Capture (preview)0

Datasets11

- Athletes
- AthletesData
- AthletesSink
- Coaches
- CoachesSink
- EntriesGender
- EntriesGenderSink
- Medals
- MedalsSink

ingestion

Athletes

AthletesSink

DelimitedText

AthletesSink

Connection

Schema

Parameters

Linked service \*AzureDataLakeStorage1Test connectionEditNewLearn more

File pathtokyo-olympic-data / raw-dataAthletes.csvBrowsePreview dataDetect for

Compression typeNo compression

Column delimiterComma (,)

Row delimiterDefault (\r,\n, or \r\n)

EncodingDefault(UTF-8)

Quote characterDouble quote (")

Escape characterBackslash (\)

Total Datasets Imported:

Microsoft Azure

Data FactoryTokyoOlympicADF9999

Search factory and documentation

azuser2356\_mml.local@techademy.com  
TECHADEMY LEARNING SOLUTIONS PRIVATE LIMITED

>>

Data Factory

Validate all

Publish all

Preview experienceOff

Factory Resources

Filter resources by name

Pipelines4

- Data\_Quality\_Checks
- Final\_Pipeline
- ingestion
- Store\_Quality\_Metrics

Change Data Capture (preview)0

Datasets11

- Athletes
- AthletesData
- AthletesSink
- Coaches
- CoachesSink
- EntriesGender
- EntriesGenderSink
- Medals
- MedalsSink
- Teams
- TeamsSink

Tokyo Olympics Databricks workspace:

Microsoft Azure

Search resources, services, and docs (G+/I)

Copilot

azuser2356\_mml.local@techademy.learning.sou...

Home > Azure Databricks >

Azure Databricks

Techademy Learning Solutions Private Limited (tec...

Create

Manage view

...

Filter for any field...

Name

↑

Codingchallengeworkspace

...

hexaworkspaceproject

...

Myworkspacehexa

...

TokyoOlympicDatabrick

...

TokyoOlympicDatabrick

Azure Databricks Service

Search

Delete

Overview

Essentials

JSON View

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Settings

Automation

Help

Status

Active

Resource group

rg-azuser2356\_mml.local-eylnK

Location

West US 2

Subscription

MML Learners

Subscription ID

2a3c6418-97b9-4d96-a24b-2c2d7633d375

Tags (edit)

Add tags

Managed Resource Group

databricks-rg-TokyoOlympicDatabrick-nyrdz6z5j4y6w

URL

https://adb-35562861446958222.azuredatabricks.net

Pricing Tier

Standard (Apache Spark - Secure with Microsoft Entra ID) (Click to cha...

Launch Workspace

Databricks notebooks:

Microsoft Azure

databricks

Search data, notebooks, recents, and more...

CTRL + P

TokyoOlympicDatabrick

A

New

Workspace

Recents

Catalog

Workflows

Compute

Data Engineering

Job Runs

Workspace

Home

Workspace

Repos (Legacy)

Favorites

Trash

Workspace > Users >

azuser2356\_mml.local@techademy.com

Share

Create

Name	Type	Owner	Created at
Data_Quality_Checks	Notebook	azuser2356_mml.local	12/19/2024, 12:05:59 AM
Store_Quality_Metrics	Notebook	azuser2356_mml.local	12/19/2024, 12:06:11 AM

Cluster Creation:

Microsoft Azure

databricks

Search data, notebooks, recents, and more...

CTRL + P

TokyoOlympicDatabrick

A

New

Workspace

Recents

Catalog

Workflows

Compute

Data Engineering

Job Runs

Machine Learning

Playground

Experiments

Features

Models

Serving

Partner Connect

Compute

azuser2356\_mml.local's Cluster

Terminate

Edit

Configuration

Notebooks (0)

Libraries

Event log

Spark UI

Driver logs

Metrics

Apps

Spark compute UI - Master

Multi node

Single node

Access mode

Single user access

Single user

azuser2356\_mml.local

Performance

Databricks Runtime Version

15.4 LTS (includes Apache Spark 3.5.0, Scala 2.12)

Use Photon Acceleration

Node type

Standard\_D4ads\_v5

16 GB Memory, 4 Cores

Terminate after

120

minutes of inactivity

Tags

No custom tags

Automatically added tags

Summary

1 Driver

16 GB Memory, 4 Cores

Runtime

15.4.x-scala2.12

Photon

Standard\_D4ads\_v5

2 DBU/h

Linked Services in ADF:

Microsoft Azure | Data Factory | TokyoOlympicADF9999

Search factory and documentation

azuser2356.mml.local@techademy.com  
TECHADEMY LEARNING SOLUTIONS PRIVATE LIMITED

» Data Factory | Validate all | Publish all

Preview experience | Off

General

Factory settings

Connections

Linked services

Integration runtimes

Microsoft Purview

Source control

Git configuration

ARM template

Author

Triggers

Global parameters

Data flow libraries

Security

Credentials

Linked services

Linked service defines the connection information to a data store or compute. [Learn more](#)

+ New

Filter by name | Annotations : Any

Showing 1 - 7 of 7 items

Name	Type	Related	Annotations
AzureDatabricks1	Azure Databricks	2	
AzureDataLakeStorage1	Azure Data Lake Storage Gen2	6	
HttpAthletes	HTTP	1	
HttpCoaches	HTTP	1	
HttpMedals	HTTP	1	
HttpServer1	HTTP	1	
HttpTeams	HTTP	1	

Data Quality check pipeline:

Microsoft Azure | Data Factory | TokyoOlympicADF9999

Search factory and documentation

azuser2356.mml.local@techademy.com  
TECHADEMY LEARNING SOLUTIONS PRIVATE LIMITED

» Data Factory | Validate all | Publish all

Preview experience | Off

Factory Resources

Filter resources by name

Pipelines | 4

Data\_Quality\_Checks

Final\_Pipeline

ingestion

Store\_Quality\_Metrics

Change Data Capture (preview) | 0

Datasets | 11

Athletes

AthletesData

AthletesSink

Coaches

CoachesSink

EntriesGender

EntriesGenderSink

Medals

MedalsSink

Activities

Search activities

Move and transform

Synapse

Azure Data Explorer

Azure Function

Batch Service

Databricks

Data Lake Analytics

General

HDInsight

Iteration & conditionals

Machine Learning

Power Query

Validate | Debug | Add trigger

Notebook

Notebook1

General | Azure Databricks | Settings | User properties

Databricks linked service \* | AzureDatabricks1 | Test connection | Edit | New

Connection successful

Store Quality metrics Pipeline:

Microsoft Azure

Data FactoryTokyoOlympicADF9999

Search factory and documentation

azuser2356.mml.local@techademy.comTECHADEMY LEARNING SOLUTIONS PRIVATE LIMITED

Validate allPublish all

Preview experienceOff

Factory Resources

Filter resources by name

Pipelines4

Data\_Quality\_Checks

Final\_Pipeline

ingestion

Store\_Quality\_Metrics

Change Data Capture (preview)0

Datasets11

Athletes

AthletesData

AthletesSink

Coaches

CoachesSink

EntriesGender

EntriesGenderSink

Medals

MedalsSink

Teams

Activities

Search activities

Move and transform

Synapse

Azure Data Explorer

Azure Function

Batch Service

Databricks

Data Lake Analytics

General

HDInsight

Iteration & conditionals

Machine Learning

Power Query

ValidateDebugAdd trigger

Notebook

Notebook2

GeneralAzure DatabricksSettingsUser properties

Databricks linked service \* AzureDatabricks1 Test connection Edit + New

Final Execution Pipeline:

Microsoft Azure

Data FactoryTokyoOlympicADF9999

Search factory and documentation

azuser2356.mml.local@techademy.comTECHADEMY LEARNING SOLUTIONS PRIVATE LIMITED

Validate allPublish all

Preview experience

Factory Resources

Filter resources by name

Pipelines4

Data\_Quality\_Checks

Final\_Pipeline

ingestion

Store\_Quality\_Metrics

Change Data Capture (preview)0

Datasets11

Data flows0

Activities

Search activities

Move and transform

Synapse

Azure Data Explorer

Azure Function

Batch Service

Databricks

Data Lake Analytics

ValidateDebugAdd trigger

Execute Pipeline ingestion

Execute Pipeline Data\_Quality\_Che... Data\_Quality\_Checks

Execute Pipeline Store\_Quality\_Met... Store\_Quality\_Metrics

Alerts and Metrics:

Microsoft Azure

Data FactoryTokyoOlympicADF9999

Search factory and documentation

azuser2356.mml.local@techademy.comTECHADEMY LEARNING SOLUTIONS PRIVATE LIMITED

Alerts & metrics

RefreshMetricsNew alert rule

ALERT	ENABLED	RESOURCE TYPE	RESOURCES	ACTIONS
NewAlert	On	Pipeline	0	

Target Criteria:

Microsoft Azure

Data FactoryTokyoOlympicADP9999

Search factory and documentation

azuser2356\_mml.local@techademy.com

TECHADEMY LEARNING SOLUTIONS PRIVATE LIMITED

»

Alerts & metrics

«

RefreshMetricsNew alert rule

ALERT	ENABLED	RESOURCE TYPE
NewAlert	On	Pipeline

Alert rule name \*

NewAlert

Description

Severity \*

Sev0

Target criteria

Actions

Whenever Pipeline Failed Runs metric is Greater

+ Add criteria

Configured Email notification:

Notifications

Action group type

Actions

ActionNotification

1 Email

+ Configure notification

Enable rule upon creation

On

Submitted by:  
Aathirainathan P